J. Clin. Chem. Clin. Biochem.
Vol. 21, 1983, pp. 813–821

# Results of Quality Control Surveys of Radioimmunological Determinations of Thyrotropin in Newborns

By *G. Röhle*

*Institut für Klinische Biochemie der Universität Bonn,*

*U. Voigt*

*Institut für Medizinische Statistik, Dokumentation und Datenverarbeitung der Universität Bonn,*

*R. Kruse*

*Institut für Klinische Biochemie der Universität Bonn* und

*T. Torresani*

*Universitäts-Kinderklinik Zürich*

(Received March 12/July 4, 1983)

**Summary:** Within the quality control scheme of the Deutsche Gesellschaft für Klinische Chemie, seven quality control surveys of thyrotropin (TSH) determinations in blood dried on filter paper have been carried out since 1980. Ninety-six screening laboratories from 12 European countries took part in these surveys. In a single survey each participant usually analysed four different samples; each of these consisted of three spots of dried blood spiked with defined amounts of thyrotropin. For the evaluations of the surveys the participants were asked to give information about their analytical results, and from these, their diagnostic classifications. The medians of the analytical results correlated well with the given thyrotropin concentrations, but the individual estimations from different laboratories varied greatly. Major discrepancies of classification were also apparent, both in the given thyrotropin concentrations and in the individual estimations.

Two special collaborative studies with nine selected laboratories showed on the one hand that analysis of the largest possible part of the dried blood sample can help to optimize the precision of the results; on the other hand, considerable drawbacks related to the reagents and the methods were sometimes observed.

*Ergebnisse aus Ringversuchen für radioimmunologische Thyrotropinbestimmungen bei Neugeborenen*

**Zusammenfassung:** Innerhalb des Systems der Externen Qualitätskontrolle der Deutschen Gesellschaft für Klinische Chemie wurden seit 1980 sieben Ringversuche für Bestimmungen von Thyrotropin (TSH) in Blut, das auf Filterpapier getrocknet ist, durchgeführt. An diesen Ringversuchen beteiligten sich 96 Screening-Laboratorien aus 12 europäischen Ländern. In jedem Ringversuch konnten die Teilnehmer durchschnittlich vier verschiedene Proben untersuchen, die jeweils aus einer Filterpapierkarte mit drei Tropfen getrockneten Blutes bestanden, das mit definierten Mengen von Thyrotropin versetzt war. Für die Auswertungen gaben die teilnehmenden Laboratorien neben ihren Analysenergebnissen die daraus folgenden individuellen diagnostischen Beurteilungen an. Während die Mediane der Analysenergebnisse gut mit den vorgegebenen Thyrotropinkonzentrationen übereinstimmten, war die Variabilität der individuellen Ergebnisse aus verschiedenen Laboratorien sehr groß. Auch in den diagnostischen Beurteilungen ergaben sich erhebliche Diskrepanzen; sowohl bei Zugrundelegung der vorgegebenen Konzentrationen als auch auf der Basis der individuellen Ergebnisse.

Zwei gesonderte Studien, an denen sich neun ausgewählte Laboratorien beteiligten, zeigten einerseits, daß die Verwendung eines möglichst großen Teils des Probenmaterials für die Analysen zu einer Optimierung der Präzision der Ergebnisse beitragen kann, andererseits ließen sie zum Teil erhebliche Mängel der angewendeten Reagenzien und Methoden erkennen.

## Introduction

According to the data (1) available at present, congenital hypothyroidism shows an average frequency of occurrence of one case in every 3,500 births. For the early diagnosis of this congenital defect the determination of thyrotropin (TSH) by radioimmunoassay in the blood of newborns is widely accepted. The preferred sample material is blood taken from the heel of the newborn on the fourth or fifth day of life, applied to filter paper and dried. In the analytical laboratory defined fields of the blood-stained filter paper are punched out and the radioimmunoassay of thyrotropin is performed after elution of the punch.

From the beginning the kind of support material used for the specimens, which is unusual in clinical chemical analytics, caused certain doubts about the reliability of the analytical results. The classification of the testing principle as a semi-quantitative screening method seemed to be justified (2, 3). On the other hand, it is conceivable that a false negative result would result in delayed therapy, with irreversible neurologic defects in the child as the possible consequence. Analysts in charge were therefore very concerned to examine the reliability of the analytical results by means of external quality control. On the basis of this interest, quality control surveys for determinations of thyrotropin from dried blood were introduced into the external quality control scheme of the Deutsche Gesellschaft für Klinische Chemie. After a pilot study in May 1980, seven quality control surveys with Professor Ruth Illig, Zürich, as consultant, have now been carried out with the participation of laboratories from 12 European countries. The evaluations comprised the values of the analytical results as well as their diagnostic classification by the participants.

Two separate quality control surveys were designed to detect possible sources of error. The first quality control survey examined whether the size of the punched sample and the precision of the results were interdependent within one series of analyses. The second quality control survey was aimed at investigating how far the working standards of some commercial kits corresponded to the reference preparation MRC 68/38.

## Materials and Methods

### Control samples

The samples for the quality control surveys 1/80, 1/81 and 2/81 were provided by Deutsche Pharmacia, Freiburg, and Henning, Berlin. For the next quality control surveys the samples were prepared in the laboratory of our institute. Filter paper from Schleicher & Schüll, Dassel, product No. 2992, served as support material. The specimens themselves were produced from human EDTA blood with a basic thyrotropin concentration of 1 mU/l, and in each case mixed with defined amounts of the international reference preparation MRC 68/38 (NIBSC-London). This material was applied to filter paper in portions of 50 µl and then dried. One sample consisted of filter paper with three spots of blood. The concentrations of thyrotropin resulting from the preparation are given in table 1.

### Quality control surveys

Seven quality control surveys were carried out in the following months: October 1980 (1/80), February 1981 (1/81), September 1981 (2/81), December 1981 (3/81), February 1982 (1/82), May 1982 (2/82) and November 1982 (3/82). The number of different samples for analysis by each participant in each quality control survey is shown in table 1. For the analysis of the samples the participating laboratories in Germany had one week, the participants from other European countries two weeks because of the longer mailing time.

### Documentation of the results and findings

Together with control samples, the participants in the quality control surveys received a registration form. The results of two single determinations and their mean in mU per l of blood were entered on these forms. In addition, a diagnostic classification of the analysis result was required:

(1) normal

(2) a case of congenital hypothyroidism possibly or

(3) probably pathologic.

Participants were also asked what further measures were taken in consequence of the first analysis result:

(1) None?

(2) Analysis of the third spot of blood?

(3) The ordering of a new sample card?

(4) The ordering of a serum sample?

As further information, the name of the kit producer was requested, provided that a commercially available test kit had been used for the determinations.

In the quality control survey 3/82 each participant was also asked about the highest concentration of thyrotropin he would consider to be normal for a 5-days-old infant, and which value of thyrotropin in routine analytics was the lowest that would result in further measures.

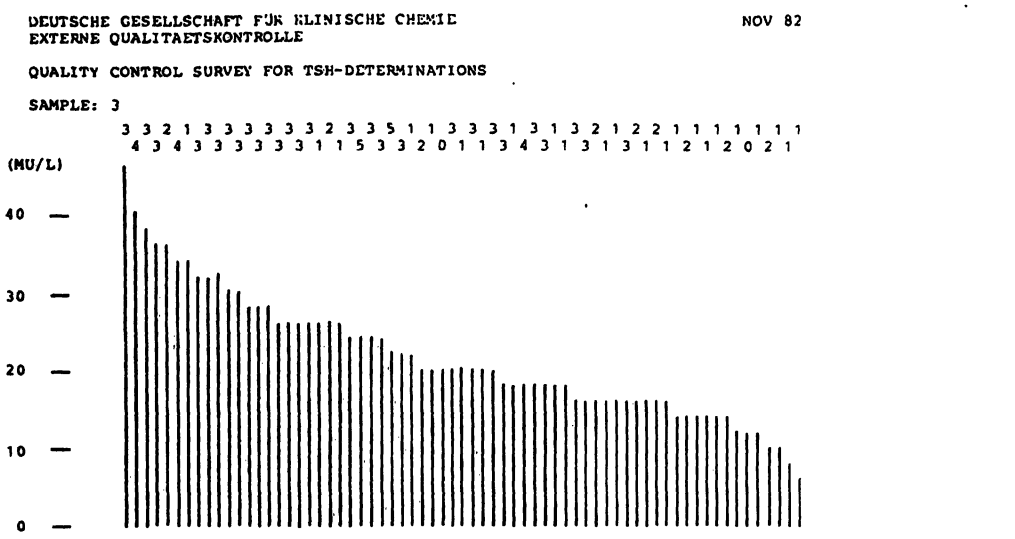


Fig. 1. Example of general evaluation of a survey for thyrotropin determinations (survey 3/82; specimen 3), including statistics on diagnostical classifications. Each vertical line represents the analytical result of an individual laboratory. MU/L = mU/l.

### Evaluation

When the survey was finished the data were collected and evaluated by means of electronic data processing. The results of each sample of the quality control survey were summarized as shown in figure 1. The height of each vertical line represents the value determined by one laboratory. The values were listed in descending order. Above the diagram the code numbers for the diagnostic classification belonging to each value were noted. The summary below the diagram took into account the statistical parameters of mean, standard deviation and coefficient of variation. The statistical values of the diagnostic classifications and of any further measures were also recorded.

*Special survey "diameter of paper discs" (May 1982)*

Nine participating laboratories – five laboratories from producers of kits and four from screening laboratories for the screening of thyrotropin – measured the concentrations of thyrotropin in two sets of samples (A = 13 mU/l; B = 31 mU/l) by means of eight-fold determinations. Each eight-fold determination was carried out with the following samples: two samples of 3 mm, one sample of 4.25 mm, one of 6 mm and, as far as technically possible, one sample with a disc diameter of 8 mm. From the values of each eight-fold determination the precision in the series was calculated as a coefficient of variation.

*Special survey "Hidden dose-response curve" (May 1982)*

In accordance with the model of quality control surveys developed by *Marschner* et al. (4), the same laboratories that had participated in the quality control survey "diameter of paper discs" also measured in duplicate the concentrations of thyrotropin in eight different blood samples dried on filter paper. Six of the samples – all of them were unknown to the participants ("hidden standards"; samples 1–6) – contained the following concentrations of thyrotropin by addition of the reference preparation MRC 68/38: 1 mU/l, 7 mU/l, 13 mU/l, 25 mU/l, 31 mU/l, and 43 mU/l. Two samples were prepared from the blood of patients (samples 7 and 8). The participants of the quality control survey returned the following data for the evaluation:

1. The given concentrations of thyrotropin and the counts per minute of the working standards of the tests.

2. The counts per minute for the unknown samples ("hidden standards" and samples from the patients) and the corresponding concentrations of thyrotropin read from the dose-response curve of the test. From the returned counts per minute, a dose-response curve on the basis of the working standards and a "hidden" dose-response curve on the basis of the samples 1–6 were calculated for each participant using a standardized procedure. From the counts per minute for samples 7 and 8 and the individual hidden dose-response curve, the resulting concentrations of thyrotropin were computed and set against the results achieved by the participants.

## Results and Discussion

96 laboratories from 12 European countries participated in the 7 quality control surveys. Most of them used commercial test combinations for the determinations of thyrotropin, but the proportion of laboratories that used their own methods was – at 25% – relatively large (5).

## Within-series precision

Because of the unusual nature of the support material, the question arose as to the reproducibility of the results under the most favourable conditions. The smallest differences among the individual results caused by systematic errors were to be expected within a series of analyses. From the results of duplicate determinations, the mean within-series imprecision could be calculated. In figure 2 this imprecision, calculated as coefficient of variation, was set against the median of the values for each analysed sample.

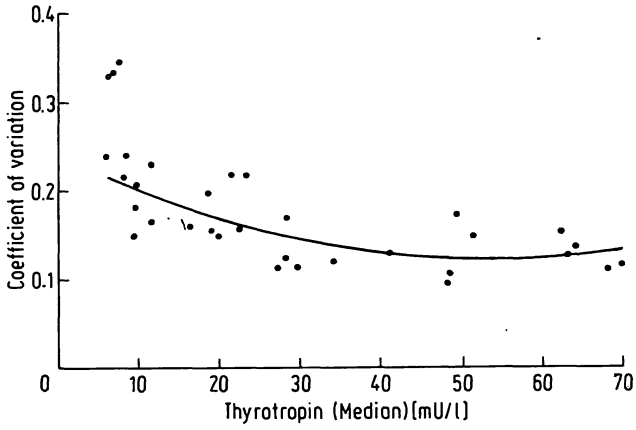


Fig. 2. Mean intra-assay coefficients of variation calculated from the values of double determinations performed by the survey participants versus the median of all analytical results for thyrotropin of a given specimen.

Below a thyrotropin concentration of 10 mU/l, coefficients of variation up to 40% were observed. In the most frequent decision area – at 20 mU/l – they were between 15% and 20%. Under the best conditions, coefficients of variation of about 10% were found. Apart from inevitable random errors by the analysts, the following factors may have affected the within-series precision: heterogeneity of the support material (filter paper), and different concentrations of thyrotropin in the dried blood spot due to a chromatographic effect (6).

## Variability of the results from different laboratories

Differing analytical conditions in the individual laboratories had a major influence on the results of the quality control surveys, giving rise to considerable variation in the results from different laboratories. The parameter "coefficient of variation" proved to be unsuitable for the description of the distribution, as each group of values did not follow a normal distribution pattern. Instead, in figure 3 the values of the 84% percentiles and of the 16% percentiles relative to the corresponding median were noted; if the values were distributed normally, the mapping points would be in symmetric order to the abscissa and their distance from the abscissa one standard deviation in each direction.

It might be supposed that the considerable variability of the results from different laboratories was due to the differing quality of the commercial kits. This would be in accordance with the results of a comparison of commercial kits performed by *van Thiel* et al. (7) in 1980.

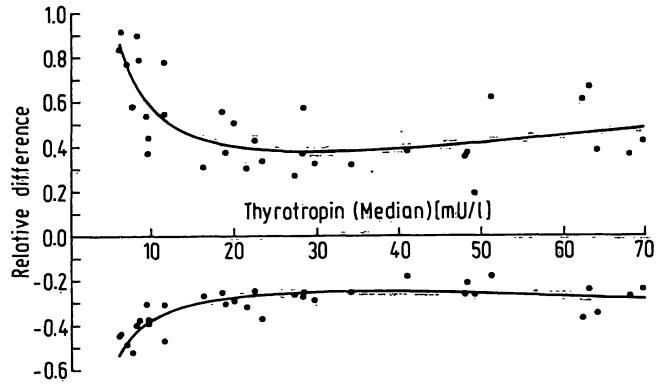


Fig. 3. Relative differences between median (0.0) and 16% (−) and 84% (+) percentiles, respectively, of the analytical results of thyrotropin determinations for a given specimen.

Figure 4 shows the position of the medians, as well as the level of the minimum and maximum values, for the results from laboratories using commercial kits for sample 4 in the quality control survey 3/82. This example confirmed the results of *van Thiel* et al. (7)

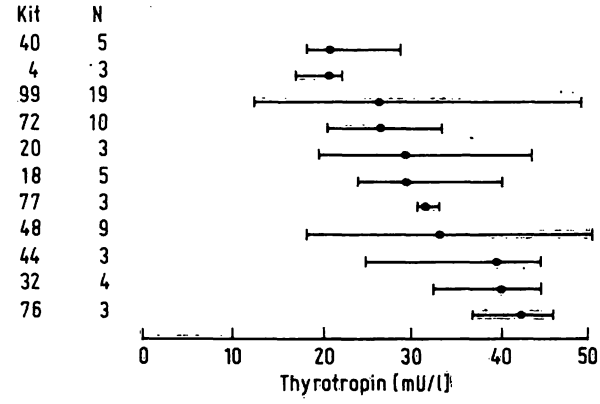


Fig. 4. Median and range of the analytical results for thyrotropin (survey 3/82; specimen 4) with respect to the kit used.

as can be seen from the varying positions of the medians; but a direct comparison of these results of an intralaboratory study with the interlaboratory imprecision of external quality control surveys was not possible because of the different experimental conditions. The results of one of the special surveys described below, lead to the conclusion that at least in some cases insufficient coincidence of the working standards with the reference preparation MRC 68/38 was the reason for incorrect results. On the other hand, the small number of results in the sub-groups was not a solid basis for further conclusions. However, it might be assumed that procedural errors in the individual laboratories contributed considerably to the substantial variability among the laboratories.

Tab. 1. Thyrotropin concentrations (mU/l) of the specimens analysed during seven quality control surveys.

| Specimen | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| Survey | | | | | | |
| 1/80 | * | * | * | – | – | – |
| 1/81 | * | * | * | * | – | – |
| 2/81 | * | * | * | * | * | 31 |
| 3/81 | 8.5 | 23.5 | 53.5 | 8.5 | – | – |
| 1/82 | 11 | 38.5 | 19 | 61 | – | – |
| 2/82 | 17.5 | 8.5 | 31 | 46 | – | – |
| 3/82 | 9.5 | 53.5 | 23.5 | 31 | – | – |

* Thyrotropin concentration not defined

## Accuracy

For most of the samples employed in the quality control surveys, the exact concentrations of thyrotropin were given with good approximation (tab. 1). Though some of the individual results – as reported above – differed considerably from these target values, the medians of the results in nearly all cases showed a good correlation with the target values for the different samples (fig. 5). Thus this analytical principle for the determination of thyrotropin in dried blood produces exact results under the conditions of quality control surveys. However, the wide variance of the results from different laboratories showed that the precision of results was insufficient, and that a reduction of the systematic and random errors should be attempted.

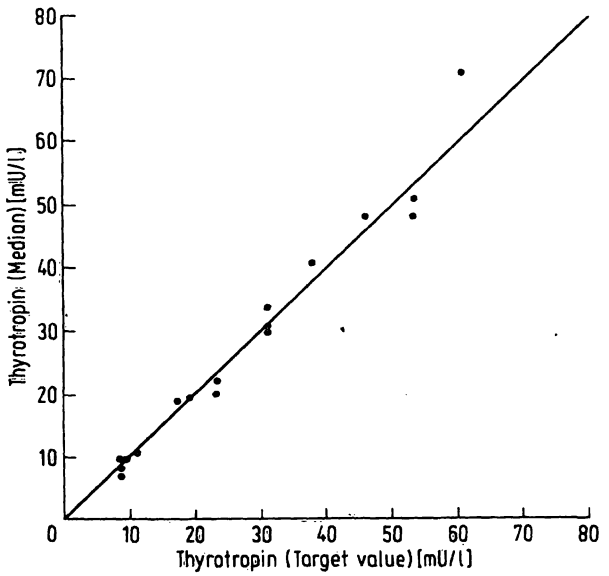


Fig. 5. Correlation between the given thyrotropin concentrations of the survey specimens and the medians of participants' results.

## Diagnostic classification

Whenever a complex principle of analysis is used, as for the determination of thyrotropin, it cannot be expected that the results from different laboratories will be fully comparable, especially when the method is new. In spite of this, compensation of interlaboratory differences in reference ranges could finally lead to a unified diagnostic classification of the analysis results. A request during the quality control survey 3/82 for the intra-laboratory upper limit of the reference range, and for the lower decision limit that would lead to further diagnostic measures, made clear that these limits were being defined quite differently. The region in which the individual laboratories set the upper limit of the reference range varied from 8 mU/l to 50 mU/l (fig. 6a). A similar degree of variation was found in the limits – based on the results of the routine analyses – beyond which further steps would have been taken, namely from 8 mU/l to 35 mU/l (fig. 6b).

Though the distribution of the values for the decision limits and the distribution of the analysis results for a sample (the concentration of thyrotropin which was within the decision range (fig. 1)) were comparable, the combination of both distributions in a given sample definitely did not lead to a unified diagnostic classification in practice. In figure 7 the relative proportion classified as "normal" was correlated with the median of the corresponding analysis results for all the samples analysed so far in the quality control surveys (cf. fig. 5, showing that median and correct value largely coincided). Obviously the variability of the analysis results, as well as a general uncertainty as to the optimal position of the decision limit, were the reason for the discrepancies observed with the classification of a given sample. From the present data it cannot be stated how often unnecessary subse-

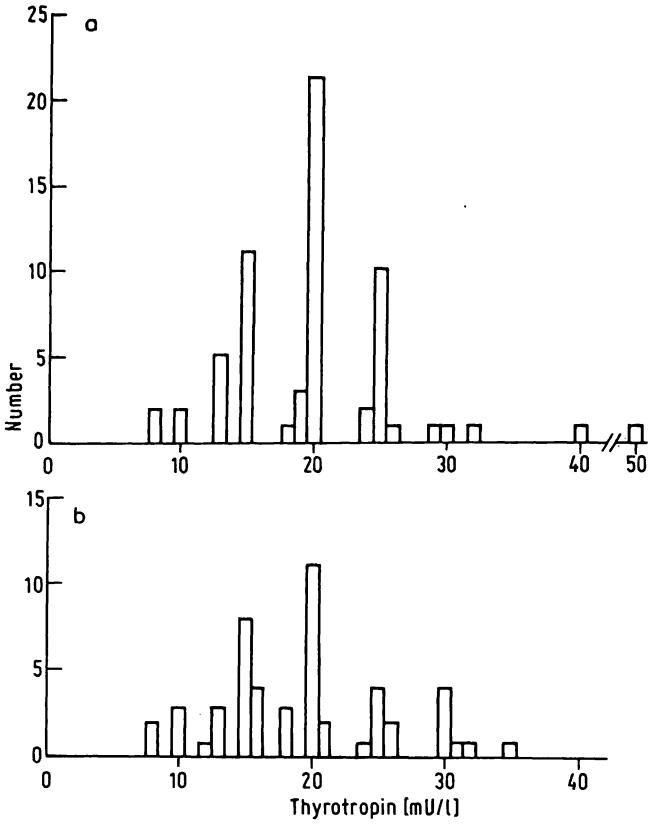


Fig. 6. Classification and decision limits on the basis of individual analytical results. Result of an inquiry.
a) Frequencies of the positions of the highest thyrotropin concentration in blood that is considered normal by individual laboratories (63) in the case of a five-days-old infant.
b) Frequencies of the positions of the lowest thyrotropin concentration in blood that calls for further measures in individual laboratories (53) in the case of a five-days-old infant.

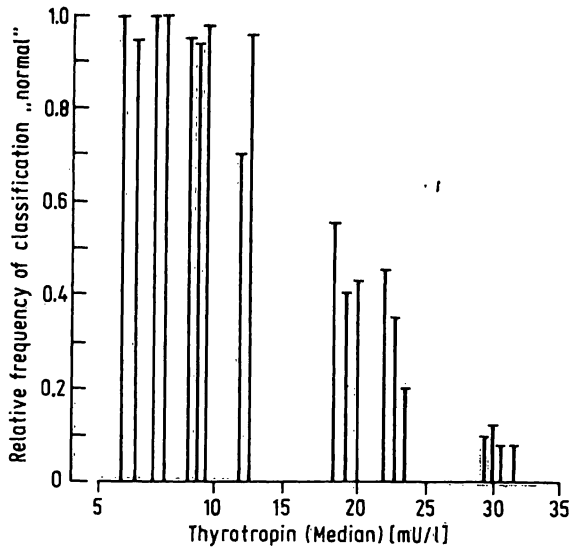


Fig. 7. Relative frequency of the diagnostic classification "normal" versus the medians of the results of thyrotropin determinations for given survey specimens.

quent determinations were caused by laboratories with a very low decision limit; and whether laboratories with a very high decision limit missed some cases of congenital hypothyroidism. However, it is very probable, on the basis of these results, that at present the number of false classifications is significantly above an acceptable rate of error. It is to be hoped that the intensive exchange of experience that takes place among the European screening laboratories soon leads to an approximation of the decision criteria.

## Special surveys

Of the possible error sources that may have affected the reliability of the analysis results, two were considered further in collaborative studies:

– The dependence of the within-series imprecision on the size of the disc.

– The comparability of the kit standards with the reference preparation MRC 68/38.

As *Hoepfner* (6) showed, the concentrations of thyrotropin in a spot of blood dried on filter paper may be considerably higher at the border of the spot than in the middle, because of a chromatographic effect. Consequently, the paper discs should be as large as possible so as to eliminate these differences of concentration to improve the within-series precision. At present, discs with diameters from 3 mm to 8 mm are used for routine analyses. In order to determine which diameters of discs are best suited to achieve good within-series precision, nine laboratories determined, in two sets of the same sample material, the concentrations of thyrotropin by means of eight-fold determinations. Each eight-fold determination was performed with the following diameters of discs: 3 mm (two), 4.25 mm, 6 mm, and 8 mm. The coefficients of variation calculated from each series of analysis are presented in table 2. Discs with a diameter of 8 mm were not suitable for many of the methods used, and four participants in this study could not obtain any results for them. Therefore this size of sample cannot be judged in this context. The other sizes showed, however, – in spite of the differing data of the individual laboratories – that for both concentrations the best precision was achieved by discs with a diameter of 6 mm. Similarly, *van Thiel* et al. (7) compared discs of 6.2 mm and 8 mm diameter respectively and with the larger disc they observed a better sensitivity but no improvement of precision.

Tab. 2. Coefficients of variation of the intra-assay precision of eight-fold thyrotropin determinations achieved by nine laboratories using varying sizes of discs.

| | Specimen A (13 mU/l thyrotropin) | | | | Specimen B (31 mU/l thyrotropin) | | | |
|---|---|---|---|---|---|---|---|---|
| Diameter (mm) of discs | 3 (two) | 4.25 | 6 | 8 | 3 (two) | 4.25 | 6 | 8 |
| Laboratory | | | | | | | | |
| 1 | 0.28 | – | 0.18 | 0.20 | 0.17 | 0.36 | 0.04 | 0.10 |
| 2 | 0.08 | 0.39 | 0.12 | 0.07 | 0.06 | 0.15 | 0.11 | 0.10 |
| 3 | 0.22 | 0.12 | 0.12 | 0.26 | 0.17 | 0.11 | 0.15 | 0.13 |
| 4 | 0.15 | 0.08 | 0.07 | – | 0.07 | 0.04 | 0.06 | – |
| 5 | 0.50 | 0.34 | 0.26 | – | 0.28 | 0.15 | 0.15 | – |
| 6 | 0.44 | 0.26 | 0.11 | – | 0.14 | 0.22 | 0.15 | – |
| 7 | 0.32 | 0.44 | 0.39 | – | 0.29 | 0.62 | 0.28 | – |
| 8 | 0.17 | 0.24 | 0.39 | 0.27* | 0.16 | 0.13 | 0.07 | 0.32** |
| 9 | 0.15 | – | 0.13 | 0.48 | 0.29 | – | 0.06 | 0.06 |
| Median | 0.22 | 0.26 | 0.13 | 0.26 | 0.17 | 0.15 | 0.11 | 0.10 |

* seven-fold determination
** four-fold determination

In another survey, the trial design of which was performed in accordance with the „Münchner Modell" (4, 8), nine laboratories received eight different samples each with a concentration of thyrotropin that was unknown to the participants.

The comparison of the dose-response curves of the working standards and the "hidden standards" resulted partly in very different combinations (fig. 8a–c). In three laboratories the coincidence of both curves was very good, as shown in figure 8a. In four laboratories a more or less distinct parallel displacement was observed, as seen in the example in figure 8b. Possibly these displacements are mainly caused by errors in the calibration of the working standards. Additional drawbacks of the method and probably handling mistakes, however, must be supposed for the case presented in figure 8c. Here the curves cross and diverge to an increasing extent in an important part of the range of measurement. A similar pattern of the curves was observed by another participating laboratory, although with opposite and smaller discrepancies.

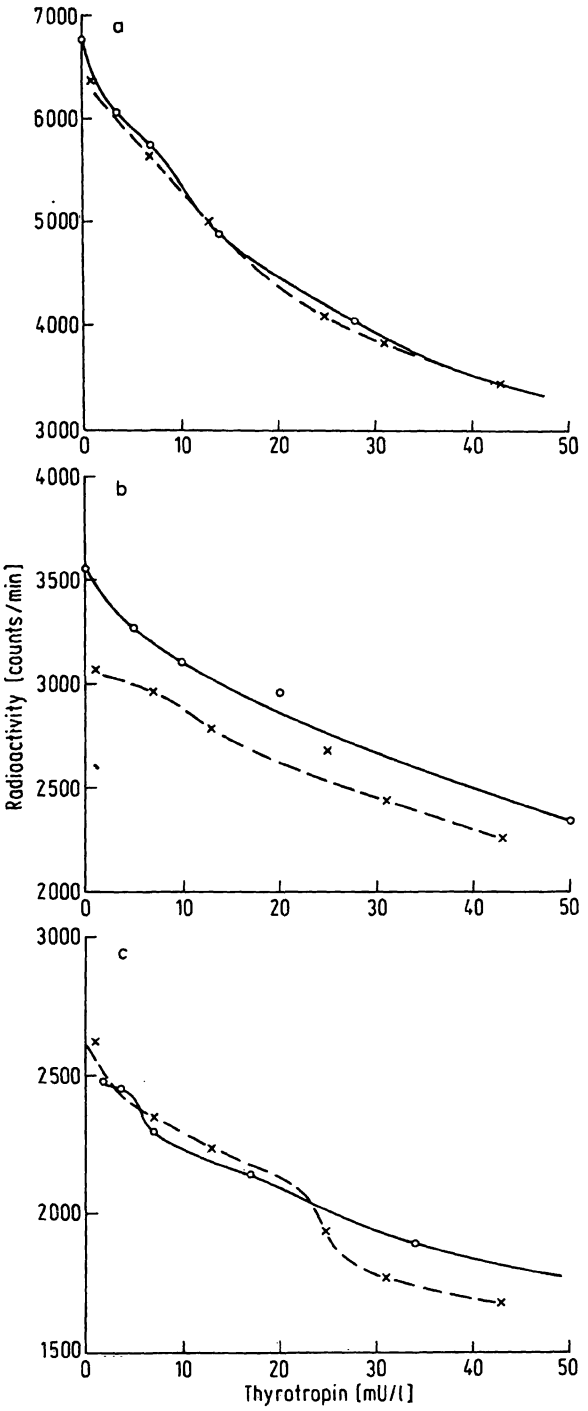


Fig. 8. Dose-response curves for thyrotropin calculated from the counts measured by three individual laboratories
——— for the working standards of the kit used
– – – – for the hidden standards among the survey specimens.

Using the counts per minute for samples 7 and 8 (samples from patients), the concentrations of thyrotropin (that should have resulted on the basis of the hidden dose-response curve from the individual laboratories) were calculated. In figures 9a–d they are compared with the analysis results that eight of the nine participating laboratories had returned (the results from laboratory 7 were unsuitable for the comparison). The expected optimization of precision throughout the laboratories for the values computed on the basis of the hidden dose-response curve was only partially attained. With sample 7 (fig. 9a and b) the range of distribution was as wide, but differently positioned. This situation was also confirmed by the standard deviations, which were 4.5 mU/l in both cases. At least for the sample 8 (fig. 9d), the range of

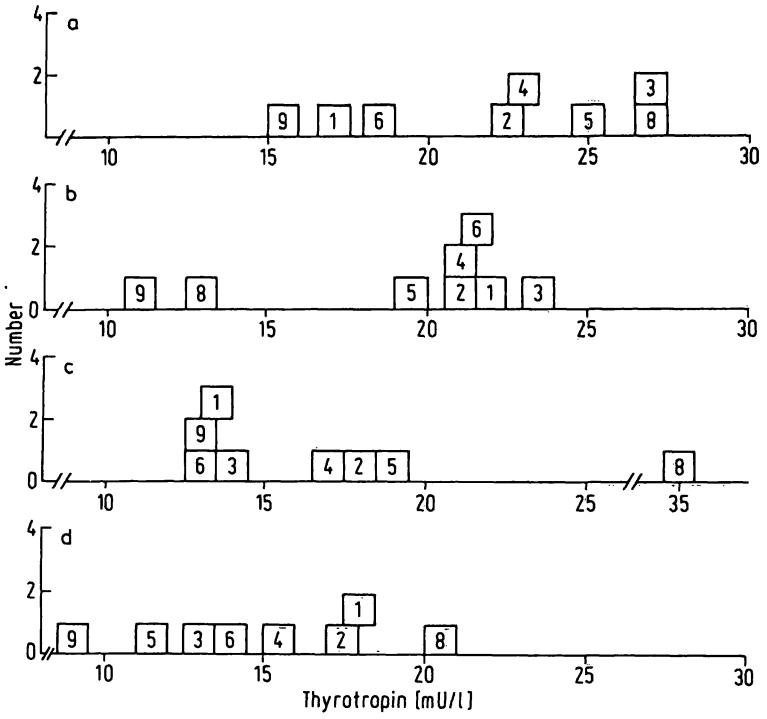


Fig. 9. Distributions of the analytical results for thyrotropin in two different specimens (a, b and c, d respectively) from hypothyreotic patients established by eight laboratories. The numbers within the squares symbolize the individual laboratories.
a) and c) analytical values calculated by the participants themselves.
b) and d) analytical values calculated from the hidden dose-response curves.

distribution of the values on the basis of the hidden dose-response curves was considerably smaller than for the values from the laboratories. The value from laboratory 8, which can be called an outlayer, came closer to the average of the other values after the recalculation, and the standard deviations decreased from 7.3 mU/l to 3.8 mU/l.

## Conclusion

The determination of thyrotropin in blood dried on filter paper has proved to be a very helpful method for the early diagnosis of congenital hypothyroidism. It is possible to determine the correct concentrations of thyrotropin by this method. This is evident from the satisfactory accuracy of the medians of the results in quality control surveys, as well as in the considerable reliability of the analysis values of some individual laboratories.

By use of a semiquantitative analytical method the insufficient precision of the results from different laboratories can only be improved to a certain degree. The investigations of *Höpfner* (6) indicated, however, some initial steps for a possible optimization, especially with respect to the kind of filter paper used and the size of the samples. The results confirmed that samples of the largest possible size may improve the reliability of the results. Furthermore, it could be shown that often errors in calibration were probably behind incorrect results.

For the important diagnostic classification of the analysis results, the quality control surveys revealed great differences in the location of the decision limits, taking into consideration the correct concentrations of thyrotropin as well as the intralaboratory analysis results. These uncertainties in the diagnostic classification may result partly because in some countries the screening of hypothyroidism by thyrotropin assay has been in use for only few years, so that experience is still limited. But a slight improvement in the reliability of analytic procedures – one that could be gained by taking the measures discussed above – would represent considerable progress toward exact diagnosis.

## References


1. Delange, F., Illig, R., Rochiccioli, P. & Brock-Jacobsen, B. (1981) Acta Paediatr. Scand. *70*, 1–2.
2. Illig, R. & Rodriguez de Vera Roda, C. (1976) Schweiz. Med. Wochenschr. *106*, 1676–1681.
3. Illig, R., Torresani, T. & Sobradillo, B. (1977) Helv. Paediat. Acta *32*, 289–297.
4. Marschner, I., Erhardt, E. W. & Scriba, P. C. (1976) J. Clin. Chem. Clin. Biochem. *14*, 345–351.
5. Voigt, U., Röhle, G., Kruse, R. & Breuer, H. (1982) In: Radioimmunoassay and Related Procedures in Medicine, IAEA-Proceedings, Vienna 1982, p. 607–614.
6. Höpfner, B. (1982) J. Clin. Chem. Clin. Biochem. *20*, 915–920.
7. van Thiel, Dagmar, Marschner, I., Wood, W. G., Habermann, J. & Scriba, P. C. (1980) J. Clin. Chem. Clin. Biochem. *18*, 807–816.
8. Marschner, I., Wood, W. G., van Thiel, Dagmar, Habermann, J., König, A. & Scriba, P. C. (1983) J. Clin. Chem. Clin. Biochem. *21*, 301–311.


Dr. G. Röhle
Institut für Klinische Biochemie
Sigmund-Freund-Straße 25
D-5300 Bonn 1